

## **SIMULTANEOUS MULTI-USER REAL-TIME VOICE RECOGNITION SYSTEM**

### **CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims the benefit under 35 U.S.C. 119(e) of U.S. Provisional Application No. 60/214,504 filed on June 28, 2000.

### **STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT**

[0002] This invention has been created without the sponsorship or funding of any federally sponsored research or development program.

### **BACKGROUND OF THE INVENTION**

#### **Field of the Invention**

[0003] This writing explains a method to solve the problems in generating a Multi-User Conversational Voice Log or (MVL). There are many problems and sub-problems that need to be solved in order to create a MVL. These include:

- Real-time voice recognition and capture of many people
- Distinguishing each person in a group individually
- Creating the individual voice log
- Integration of each person's voice log into a combined MVL
- Organization of the many voice logs in the proper order
- Acceptable accuracy to make the log useful
- Making the text log easily accessible or printable on request
- Having a command set that can address the control of creating a MVL

## Description of the Prior Art

### The Problem

[0004] Consider a meeting with several people in a conference room. Typically, capturing the history of the meeting is handwritten notes or minutes being taken and converted into text by a human. This requires either a non-participant of the meeting to capture notes, or a person engaged in the meeting to be the note taker and provide output. In either case, it is a burden to some person. If multiple languages are needed, people with additional skills must be used.

[0005] These and other difficulties experienced with the prior art devices have been obviated in a novel manner by the present invention.

[0006] It is, therefore, an outstanding object of the present invention to provide an effective way to create a textual representation of the discussion by multiple speakers.

[0007] It is a further object of the invention to provide a Simultaneous Multi-User Real-time Voice Recognition System and text creator that is capable of being manufactured of high quality and at a low cost, and which is capable of providing a long and useful life with a minimum of maintenance. With these and other objects in view, as will be apparent to those skilled in the art, the invention resides in the combination of parts set forth in the specification and covered by the claims appended hereto, it being understood that changes in the precise embodiment of the invention herein disclosed may be made within the scope of what is claimed without departing from the spirit of the invention.

## **A NEW METHOD FOR CAPTURING HISTORY OF A MEETING OR GROUP OF PEOPLE**

A method of fixing the problem would be to use a Conference To Text System (CTTS)

[0008] Using voice recognition in the meeting environment combined with the ability to capture every person's conversation individually, including all people in total, allows minutes to be captured real-time and converted to the format of text. Thus creating a Multi-User Conversational Voice Log or "MVL". This concept can be used in many applications spanning from a single person's conversation to a meeting of the United States House of Representatives, and everything in between. Other features can be added to such a device, for example, real time language translation by displaying text in an alternate language from the input language. However, industry and people in general cannot take full advantage of voice recognition because of many problems that exist with the existing technology.

### **BRIEF SUMMARY OF THE INVENTION**

#### **KEY COMPONENTS NEEDED**

[0009] The following sections will discuss the Conference To Text System (CTTS), which are the hardware and software components that enable the ability to generate a Multi-user Voice Log or MVL. The invention described below addresses the following problems:

- Components of the technology that do not exist.
- Existing components and technology have not been brought together and debugged to support this aspect of voice recognition.
- Training the many systems needed to recognize each person is time consuming and not feasible.

- Lack of a command set to control creation of a MVL.

[0010] Key components that make up a CTTS include:

[0011] 1) Computer hardware with high performance that can service a person individually and collaborate in a high performance local area network environment. The hardware needs to have the power and packaging to be customer accepted. A unit containing a high-speed processor, memory, mass storage, audio input, optional display, and mouse would be used for each individual to be captured. A separate computer system functioning as a Voice Log Integrator is connected to the user units by a network (FIG. 3). An operating system and voice recognition application is used on each unit. Voice Model Mobility allows users to obviate the need for training.

[0012] 2) Time Stamp Utility (TSU) is run on the CTTS. The function of the TSU is to apply a time stamp for each group of words spoken between pauses. The TSU is triggered from an interrupt signal to process sound when the sound input frequency range and sound level, and/or sound pressure is within the parameters of the individual speaking versus the person not speaking. This information is stored as part of the voice model and moved into the CTTS using VMM or some similar utility or means.

[0013] 3) After the individual voice logs are captured, a Voice Time Integrator utility organizes the voice logs into chronological order by time stamp and/or some other indexing method. If identical time stamps are encountered, it is noted on the text log.

[0014] 4) Options for text output with an index that optionally can be sent to destinations like databases, text transcripts, and audio output.

[0015] 5) Command interpreter for control and the creation of an MVL. CTTS systems need a user interface for command and control. Control can be done at the individual user level and at the group level. Additionally, other control features can be

added in the post processing stage, like specific formats, highlighted areas, other languages displayed, etc.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The character of the invention, however, may best be understood by reference to one of its structural forms, as illustrated by the accompanying drawings, in which:

[0017] FIG. 1 is a schematic diagram of a conference-to-text system (CTTS) embodying the principles of the present invention;

[0018] FIG. 2 is a schematic diagram of a single user section of a conference-to-text system (CTTS) embodying the principles of the present invention;

[0019] FIG. 3 is a schematic diagram of the hardware component diagram and connections of a conference-to-text system (CTTS) embodying the principles of the present invention;

[0020] FIG. 4 is a schematic diagram of a micro-computer packaged into a 5 ¼ inch form factor for a conference-to-text system (CTTS) embodying the principles of the present invention;

[0021] FIG. 5 shows a prototype Micro-Computer packaged in sheet metal enclosure to be mounted in a form factor of standard PC 5 ¼ inch tower enclosure slot. It is shown opened with components (3.5" disk drive removed). This system has a network port for control and connection to the CTTS integrator. It also has a USB for sound input/output. The system resides away from the user so only the microphone and microphone control is at the user location; and

[0022] FIG. 6 shows the same internal components as FIG. 5 above, configured as a prototype Micro-Computer for handheld large vocabulary voice recognition packaged

in a form factor of standard PC 5 ¼ inch slot. It is shown with the display screen that can be located at a user location in a conference room. Below the pen is the microphone connector and enable switch. It also has a network port for connecting back to the CTTS integrator. It can be used as touch screen, or with a keyboard and/or a mouse.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

### Invention descriptions

#### 1) Computer Hardware with high performance

[0023] Computer hardware to support these types of applications must include at least the following components to be effective:

- High-speed microprocessors with robust floating point features
- Large on chip and/or off chip cache
- High-capacity/fast main memory
- Quality sound input device with performance focused in the range of the human voice or signal to be translated
- An operating system specifically configured (tuned) for the application of voice recognition and data base management.

[0024] The hardware in this example is configured in packaging or enclosures that support conference room, hall, and auditorium environments. For example, each user may have a microcomputer located near a small flat screen that displays the text as it's being spoken (FIG. 6). Voice, mouse, and/or touch screen can be used to command the user level device. These miniaturized computers would connect back to a computer which functions as a Voice Log Integrator. The miniaturized computers can also be packaged in standard 5 ¼" containers that install into a 5 ¼"

computer chassis slots (FIGS. 3 and 4). Although a unique form factor for a computer, this format is common for standard PC peripherals. These 5 ¼" computers could then integrate the user components, a voice time integrator, and a data base server in one contained box. A General Voice proto-type of the 5 ¼" computer is shown in FIGS. 4 and 5. This device could also be used as a handheld transcriber.

[0025] FIG. 6 is the same proto-type unit as FIG. 5, packaged and shown running in a hand held form factor. This prototype supports a vocabulary of over 30,000 words. Results from these prototype models indicate that production models could support vocabularies with hundreds of thousands of words, including such libraries as medical and legal.

## 2) Voice Time Integrator or Dialog Integrator

[0026] The Dialog Integrator is software that executes on the CTTS system (See FIG.1). It organizes the captured voice text or voice text logs and puts them in chronological order for screen output or export to a file. In summary, there are three items to be discussed with the Dialog Integrator. These three items include 1) Time stamp, 2) Integrating many voice logs together into a Multi-User Conversation Voice Log, and 3) Taking the voice text and index for each word/sentence and putting that into a database table, text file, or some other file/format. The log file contains a time stamp or some other method to synchronize all voice logs intended for conversion to MVL. The time stamp can be done as an integrator component, or the time stamp may be placed into the log by the voice recognition software or a related utility.

[0027] To date there are no voice recognition software packages on the market that include indexing, or time stamping as the words are spoken or after a delay or pause

of some amount of time, for example. The integrator could be built into other components like SVM or VMM as well.

### 3) Index or Time stamp

[0028] The index or a time stamp is needed for the organization of log files or voice text files to be compiled in the original chronological order that occurred. This index stamp is captured real-time as the events are taking place (FIGS. 1 and 2). A real-time clock or a reference related to a specific event can be used. This index can be generated in many different ways. Here is a list of some of the ways an index can be created and used to allow voice text indexing:

#### [0029] Methods of Enabling:

- Button Activated (Press button when speaking, starts index and voice capture)
- Voice Activated (Starts index when user is speaking, stops on user stop)
- Command Activated (Voice command starts index and voice capture)

#### [0030] Methods of continuously creating index when words are spoken:

- Sound level
- Sound pressure
- Sound frequency
- Button Activated

#### [0031] Methods of indexing:

- Counter
- Clock
- Text character sequence
- Control code sequence



4) Real-time voice recognition captures software and components that maintain a reliable level of accuracy

[0032] A real-time voice recognition software package is needed to capture the voices. There are many voice recognition packages on the market and in the public domain. The voice recognition software must provide a consistent standard level of accuracy. It must also contain a very large vocabulary. The voice recognition engine may be the best place to stamp the time or index because it is closest to the source. Another important feature is a consistent and reliable level of accuracy. This is essential as the public becomes aware of how to speak to machines that recognize voice. This would allow the public to grow with the technology. The key component that could help this to work would be the use of a "Voice Accuracy Meter".

- Voice Accuracy Meter

[0033] The voice accuracy meter gives the user the ability to know when the machine will respond with the standard level of accuracy at a given point in time. The voice accuracy meter can work in many different ways. For the purpose of example, I have chosen to use a text compare approach. The text to be used for the accuracy meter could be chosen by the individual user. The user would select any text file and read a section from the text. The voice accuracy meter would then do a comparison, calculation, and feed back the results in real or in past time. It could also highlight and display the words that were incorrect, and provide a percentage or graphic output of final results.

5) Text output to destinations like ASCII text files or databases that could allow random access to any word, sentence, phrase, etc.

[0034] Output of the voice text log file is important for finding any information in the course of the meeting or spoken word. Output to straight ASCII text can be read and searched with an editor. A more powerful way of controlling searching and retrieving is by combining the voice recognition, text output, and index with a database. This allows several new features in searching and retrieving, including time based retrieval, context based retrieval, thread or concept information retrieval, and relational information, to name some of the benefits.

6) Command interpreter for controlling and creating an MVL.

[0035] In creating Multi-User Voice Logs, a user-interface and commands are needed. This new command set would address the need of turning the log on, turning the voice-capture log off, playing back logs, referencing, starting new subjects, side conversations, resetting time stamps, and deleting entries that should remain off the record. Additional commands would include a mediator's command set that allow items like "turn on all microphones" to start a meeting.

[0036] Key commands for a conference voice recognition system:

- Start meeting
- Stop meeting
- Recognize group
- Recognize user
- Pause meeting
- Print meeting
- Print individual "name"
- Index method "type"

- Strike class comment
- Start mute
- Stop mute
- Start recognize "name"
- Stop recognize "name"
- Off the record
- On the record
- Bookmark "phrase"
- Mark for correction
- List uncorrected
- List corrected
- Play voice reference
- Display user "name"

## Applications

[0037] Some of the applications that could use this technology include:

- Conferences
- Phone Calls
- Interviews
- News capturing
- Script capturing
- Hallway conversations

## Enabler of Real-time data mining

[0038] Real-time data mining can be improved using these tools. The tagging of words, phrases, concepts, and users for later reference could be accomplished using the software components as described. This allows references to be inserted while the text is being generated or spoken naturally.

## Super Voice Model (SVM)

[0039] To explain the concept of the super voice model, I will first talk about the voice model and what that means. A voice model is defined as a signal, information, or electronic data file that is a representation of a person's voice or noise. Therefore any noise that could be captured contains within it a voice model.

[0040] Normally, for voice recognition software to support large vocabularies (30,000 plus words), training the software to recognize a person's voice accurately and consistently is ongoing because of ever-changing parameters with regards to the human voice and environment. Therefore, if the hardware and software (machine) that provides recognition is not current with the parameters of the person speaking, there is a delta between the user and the machine. This delta can be the cause of and a measure of inaccuracy. As people use different machines, the delta becomes dynamic and accuracy becomes inconsistent. To allow any user the ability to unplug the "Voice Model" and plug it into the current system in use allows the user the ability to have consistent accuracy. This concept was defined in a previous patent application by Darrell Poirier, and is labeled as Voice Model Mobility (VMM).

[0041] Super Voice Models (SVM) is an extension of the voice model. The Super Voice Model, as defined by Darrell Poirier in a previous patent application, is the ability of the machine to recognize many users with a single voice model. There are

many ways to achieve a super voice model. For discussion here I will use the following example. Many voice models would be categorized using parameters that can define a group of users that need specific parameters. Then, as a person starts speaking to the machine, the real-time voice would be measured and categorized using the same parameters for that individual person. The real-time parameters would be compared and matched to one of the voice models to be used from the Super Voice Model library.

[0042] Another example of creating a Super Voice Model would be to identify and categorize individual sections of many voice models, and access them individually as the persons voice parameters are selected real-time. In other words, many voice models could be set up in something similar to a large table. Similar words, patterns, phrases, and/or other parameters would be kept in adjacent locations in the table. As the person speaks, a thread would move real-time through the table based on the parameters measured real-time from the user. This concept could also be moved directly to hardware, given the availability of the technology needed.

[0043] These examples explain in overview how a Super Voice Model could be designed or implemented, the concept being that many people use voice recognition machines with no pre-training

[0044] It is obvious that minor changes may be made in the form and construction of the invention without departing from the material spirit thereof. It is not, however, desired to confine the invention to the exact form herein shown and described, but it is desired to include all such as properly come within the scope claimed.

[0045] The invention having been thus described, what is claimed as new and desire to secure by Letters Patent is: